JOINT STOCHASTIC SAMPLING FOR RNA SECONDARY STRUCTURE PREDICTION

Arif Ozgun Harmanci[1], *Gaurav Sharma*[1,2]

[1]Dept. of Electrical and Computer Engineering,
University of Rochester,
Hopeman 204, RC Box 270126,
Rochester, NY 14627, USA
{arharman, gsharma}@ece.rochester.edu

David H. Mathews[2,3]

[2] Dept. of Biostat. and Comput. Biology
[3] Dept. of Biochemistry and Biophysics,
University of Rochester Medical Center
Rochester, NY 14642, USA
David_Mathews@urmc.rochester.edu

ABSTRACT

We present a novel method for prediction of common structures and alignment for two homologous RNA sequences. The method, termed joint sampling, is based on sampling from the structural alignment space of the sequences, i.e. the space of joint representations of common secondary structures and sequence alignments of the sequences. The structural alignment space is efficiently sampled by probabilistically generating structural alignments from building blocks termed structural alignment atoms. Structure and alignment predictions are obtained by clustering the samples of structures and alignments. The experimental results show that joint sampling offers improvements in structure prediction over a sampling method that generates structures from Boltzmann ensemble of single RNA sequence. In addition, the joint sampling offers more accurate estimate of alignment as compared to estimates from a hidden Markov model.

1. INTRODUCTION

Recent years have witnessed the discovery of several new classes of *non-coding* ribonucleic acid (ncRNA) molecules that, unlike their better understood messenger RNA (mRNA) counterparts, serve a direct functional role in cellular biology without being transcribed into proteins [1, 2]. The secondary structure of these ncRNAs, i.e. the set of base pairings between the complementary nucleotides at different positions in the linear RNA chain, mediated by hydrogen bonds, is crucial in determining their function. Computational methods for the prediction of ncRNA secondary structure are therefore of significant research interest. These methods operate on the primary structure of the RNA molecule, which is obtained as a sequence of the bases A, U, G, and C using conventional sequencing technology [1] and, using this primary structure, attempt to predict secondary structure as the set of base pairings between nucleotide pairs $(A, U), (G, C)$, and (G, U) in the RNA chain.

Evolutionary changes in an RNA molecule that substitute a base pair in its secondary structure with another feasible base pair from the possibilities outlined above, allow the secondary structure to be conserved despite a change in the primary structure, i.e. the sequence. Such changes are indeed observed in nature among homologous ncRNA sequences, which evolved from the same ancestor sequence, and offer additional information for secondary structure prediction because co-variation of nucleotides at two positions that can form a base pair indicates a base pairing that is conserved in the secondary structures of the ncRNA sequences. In fact, such comparative analysis over large sets of homologs, performed in a human effort intensive fashion, provides some of the most accurate predictions of secondary structure [3]. The success has also motivated the devel-

opment of computational methods for RNA secondary structure prediction that also operate on multiple homologous RNA sequences, which have indeed offered improvements over methods that operate on single sequences (See [4] for a review).

Stochastic sampling has recently been proposed as a promising technique for RNA secondary structure prediction using a single RNA sequence [5]. Stochastic sampling operates by sampling the set of possible secondary structures of an RNA molecule according to the (estimated) Boltzmann distribution of these structures obtained via a partition function computation [6]. A cluster analysis on the resulting structures can then provide important information about the known structure for the RNA sequence. For instance, the representative structure of the most populated cluster identified in the sample of structures can be a more accurate predictor of the structure than other single sequence prediction methods [7]. The extension of the stochastic sampling methodology to situations where multiple ncRNA homologs are available has, however, received rather limited attention. Although a heuristic iterative sampling methodology has been proposed [8], a true stochastic sampling framework has not been previously developed due to the difficulty in computing the corresponding partition function for the multiple homolog scenario.

In this paper, we propose a new method, *joint sampling*, for ncRNA secondary structure estimation based on sampling of the joint space of common structures and sequence alignments of two RNA sequences. The method is enabled by our recent work in developing a formal mathematical characterization of the joint space of common structures and alignment of two RNA sequences and the determination of the corresponding joint partition function [9, 10]. Using the joint partition function, we sample the space of common structures and sequence alignments of two RNA sequences according to the corresponding pseudo-Boltzmann distribution. This process inherently harnesses the power of comparative analysis because the sampling is constrained to the space of common structures for the sequences. Following the sampling, we perform a cluster analysis similar to [7], individually on the structures obtained for each sequence in order to obtain representative structures and the centroid of the largest cluster serves as the predicted structure. The cluster analysis is also employed on the sample of sequence alignments generated by the joint sampling for prediction of alignment of the sequences.

Our results demonstrate that the joint sampling offers more accurate predictions of secondary structure than the single sequence sampling. Furthermore, the predictions of alignment are also more accurate than those obtained with conventional hidden Markov model based sequence alignments.

2. OVERVIEW

Given two homologous RNA sequences $\mathbf{x}_1$ and $\mathbf{x}_2$, the *structural alignment* [10] $\mathcal{S}$ of the sequences is a joint representation of the common secondary structures, $\mathbf{S}_1$ and $\mathbf{S}_2$, and alignment $\mathbf{A}$ of the sequences. The joint sampling aims to generate a sample of structural alignments (each denoted by $\mathcal{S}$) from structural alignment space of the sequences, $\mathcal{S}_{\text{all}}$ as per the distribution of structural alignments in the structural alignment space, $P(\mathcal{S})$ [9, 10]. The probability of a structural alignment $\mathcal{S}$ is defined according to a pseudo-Boltzmann distribution as [9, 10]:

$$P(\mathcal{S}) = \frac{e^{-\Delta G(\mathcal{S})}}{\sum_{\mathcal{S}' \in \mathcal{S}_{\text{all}}} e^{-\Delta G(\mathcal{S}')}} \qquad (1)$$

where $\Delta G(\mathcal{S})$ is the pseudo-free energy change of the structural alignment $\mathcal{S}$ [9, 10].

Figure 1 illustrates the steps for structure prediction utilizing the joint sampling. The joint sampling generates a sample of structural alignments, denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_n$ for input sequences $\mathbf{x}_1$ and $\mathbf{x}_2$. Each structural alignment is composed of a pair of common secondary structures for the sequences and a sequence alignment. The generated sample of secondary structures for each sequence are individually clustered. In Figure 1, the clustering of structures for $\mathbf{x}_1$ identifies two clusters, $\mathbf{C}_1$ and $\mathbf{C}_2$, which are composed of single branched structures and two-branched structures, respectively. For each identified cluster, the cluster centroid is computed and the centroid of the largest cluster serves as the predicted structure with other centroids as alternative hypotheses. In Figure 1, cluster centroids for $\mathbf{x}_1$ are a single branched structure and a two-branched structure. The computation of cluster centroids are also performed for sample of alignments and for sample of structures for $\mathbf{x}_2$, which are not shown in Figure 1. The clustering process inherently identifies structural "modes" that are sufficiently distinct from each other. The method offers improvement over structure prediction by estimating and thresholding base pairing probabilities because the base pairing probability computations are averages over the distinct "modes."

3. EFFICIENT SAMPLING OF STRUCTURAL ALIGNMENT SPACE

Explicit enumeration of $P(\mathcal{S})$ over all possible structural alignments is computationally infeasible because the number of possible structural alignments is extremely large even for pairs of relatively short sequences (e.g. tRNAs). Efficient sampling of the distribution $P(\mathcal{S})$ is achieved via an iterative sampling algorithm, which builds a structural alignment from basic building blocks termed *structural alignment atoms* (SAAs) [11]. The SAAs represent the irreducible elements of structural alignments.

A structural alignment $\mathcal{S}$ can be decomposed into a set of SAAs $\{\chi(i,j,k,l)\}$, where i and j denote nucleotide indices in $\mathbf{x}_1$ with $i \leq j$, k and l denote nucleotide indices in $\mathbf{x}_2$ with $k \leq l$. For each SAA $\chi(i,j,k,l)$, $i-1$ and j are *co-incident* [10, 12] with $k-1$ and l, respectively. $\chi(i,j,k,l)$ represents one of following 11 possibilities of pairing and alignment of nucleotides at indices i, j, k, and l [11]: 1) Insertion of paired nucleotides at i and j; 2) Insertion of paired nucleotides at k and l; 3) Alignment of paired nucleotides at i and j to paired nucleotides at k and l, respectively; 4) Alignment of paired nucleotides at i and j to unpaired nucleotides at k and l, respectively; 5) Alignment of unpaired nucleotides at i and j to paired nucleotides at k and l, respectively; 6) Alignment of an unpaired nucleotide at i to an unpaired nucleotide at k; 7) Alignment of an unpaired nucleotide at j to an unpaired nucleotide at l; 8) Insertion of an unpaired nucleotide at i; 9) Insertion of an unpaired nucleotide
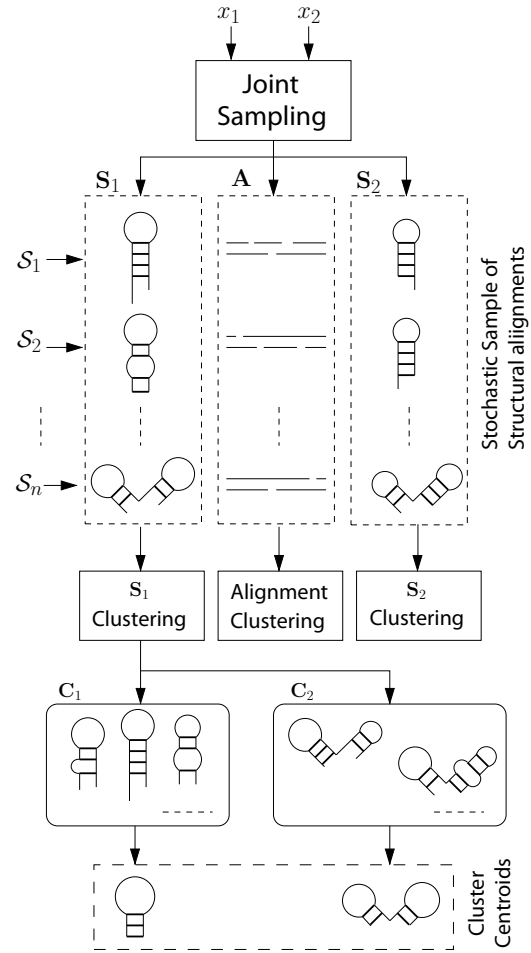

Figure 1: Overview of common structure prediction for two RNA sequences $\mathbf{x}_1$ and $\mathbf{x}_2$ by the joint sampling.

at j; 10) Insertion of an unpaired nucleotide at k; and 11) Insertion of an unpaired nucleotide at l. The SAAs represent all the possible base pairing and sequence alignment interactions between co-incident nucleotides in a structural alignment of two sequences. Any valid structural alignment can be decomposed into a set of SAAs such that the pairing and alignment of each nucleotide is identified by an SAA.

Fig. 2 illustrates decomposition of a structural alignment of two hypothetical sequences into SAAs. The dashed rectangles in Fig. 2 illustrate the decomposition of the structural alignment into 8 SAAs such that each dashed rectangle encloses the nucleotide indices whose pairing and alignment are defined by the respective SAA. The pairing and alignment of nucleotides as defined by each of 8 SAAs is indicated below:

- $\chi_1(1, 9, 1, 12)$: Paired nucleotides at 1 and 9 in $\mathbf{x}_1$ are aligned to unpaired nucleotides at 1 and 12 in $\mathbf{x}_2$, respectively.
- $\chi_2(2, 8, 2, 11)$: Paired nucleotides at 2 and 11 in $\mathbf{x}_2$ are both inserted.
- $\chi_3(2, 8, 3, 10)$: Paired nucleotides at 2 and 8 in $\mathbf{x}_1$ are aligned to paired nucleotides at 3 and 10 in $\mathbf{x}_2$, respectively
- $\chi_4(3, 7, 4, 9)$: Paired nucleotides at 3 and 7 in $\mathbf{x}_1$ are aligned to paired nucleotides at 4 and 9 in $\mathbf{x}_2$, respectively
- $\chi_5(4, 7, 5, 9)$: Unpaired nucleotide at 4 in $\mathbf{x}_1$ is aligned to unpaired nucleotide at 5 in $\mathbf{x}_2$.
- $\chi_6(5, 7, 6, 9)$: Unpaired nucleotide at 5 in $\mathbf{x}_1$ is aligned to unpaired nucleotide at 6 in $\mathbf{x}_2$.

- $\chi_7(6,7,7,9)$: Unpaired nucleotide at 6 in $\mathbf{x}_1$ is aligned to unpaired nucleotide at 7 in $\mathbf{x}_2$.
- $\chi_8(7,7,8,9)$: Unpaired nucleotide at 8 in $\mathbf{x}_2$ is inserted.

(a) Structures of sequences

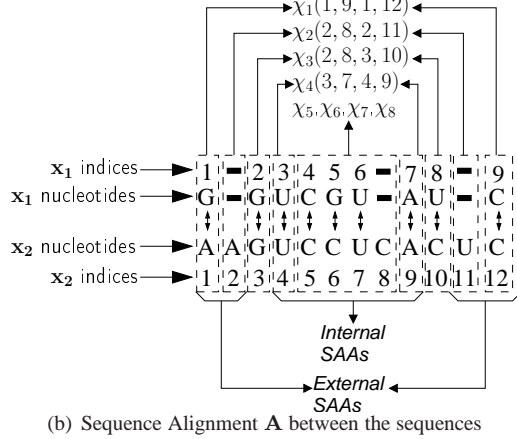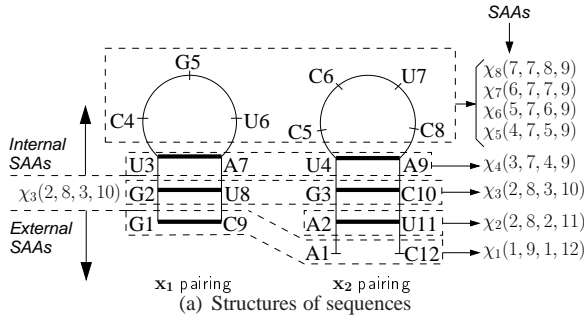(b) Sequence Alignment $\mathbf{A}$ between the sequences

Figure 2: Decomposition of a structural alignment of two hypothetical sequences into SAAs. The bold lines in Fig. 2(a) denote the base pairs. In Fig. 2(b), the aligned nucleotides are denoted by lines with double-headed arrows and a bold line in a sequence represents an insertion in other sequence.

The iterative sampling algorithm progressively builds a structural alignment by generating the current SAA, $\chi(i,j,k,l)$, via sampling according to $P(\chi(i,j,k,l)|\mathcal{S}_{\text{ext}}(i,j,k,l))$, the distribution of SAAs in pseudo-Boltzmann ensemble conditioned on the SAAs generated in the previous iterations. The SAAs generated in the previous iterations are referred to as *external* SAAs, $\mathcal{S}_{\text{ext}}(i,j,k,l)$. Therefore iterative sampling first generates the external-most SAA in the structural alignment, $\chi(1,N_1,1,N_2)$, and subsequently generates the *internal* SAAs. Given an SAA $\chi(i,j,k,l)$, the internal SAAs correspond to $\chi(i',j',k',l')$ such that $i \le i' < j' \le j, k \le k' < l' \le l$ and external SAAs correspond to $\chi(i',j',k',l')$ such that $i' \le i, j \le j', k' \le k, l \le l'$.

The steps of iterative sampling are listed in Algorithm 1. The sampling utilizes a stack that makes it possible to generate SAAs for structural alignments that contain structures with multibranch loops, i.e. bifurcated structures. Each entry in the stack stores the indices (i,j in $\mathbf{x}_1$ and k,l in $\mathbf{x}_2$) of fragments and also the set of SAAs that are generated in the previous iterations, i.e. $\mathcal{S}_{\text{ext}}(i,j,k,l)$. Initially $\{i,j,k,l,\mathcal{S}_{\text{ext}}\} = \{1,N_1,1,N_2,\emptyset\}$ is pushed on the stack. An iteration involves popping new indices and set of previously generated SAAs, i.e. $\{i,j,k,l,\mathcal{S}_{\text{ext}}\}$, from the stack, followed by a computation of conditional probability of SAAs, $P(\chi(i,j,k,l)|\mathcal{S}_{\text{ext}}(i,j,k,l))$. The conditional distribution of $\chi(i,j,k,l)$ accounts for the SAAs on structural alignments that contain single branched structures and SAAs on structural alignments that contain multibranched structures. Thus, sampling of the conditional distribution of SAAs generates one of two possibilities for the SAA at indices i,j,k,l. The first

Algorithm 1: Iterative Sampling Algorithm

possibility is that an SAA $\chi(i,j,k,l)$ on a single branched structural alignment is generated. The sequence indices whose structural alignment is established by $\chi(i,j,k,l)$ are updated such that the new values of indices point to nucleotides whose structural alignment is not yet established. The established structure and alignment of nucleotides are added to $\mathbf{S}_1$ and/or $\mathbf{S}_2$, and $\mathbf{A}$ respectively. The generated SAA is added to $\mathcal{S}_{\text{ext}}(i,j,k,l)$ and pushed onto the stack with updated indices. The second possibility is that i,j,k,l is divided into two subfragments i,ip,k,kp and $ip+1,j,kp+1,l$ such that each subfragment corresponds to new branches in structures of sequences. In this case, each subfragment of sequences is pushed onto the stack separately. SAAs for each of the subfragments are generated in subsequent iterations. The algorithm terminates when there are no more fragments whose structures and alignments are not established, i.e. when the stack is empty. The final structural alignment is formed by $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{A}$. The details of the actual implementation are available in a recent manuscript which is accepted for publication in *Nucleic Acids Research* [11].

4. CLUSTERING SAMPLES OF STRUCTURES AND ALIGNMENTS

Given a sample of structural alignments generated by the iterative sampling algorithm, the sample sets of common secondary structures of each sequence and sample set of sequence alignments are clustered individually using the *diana* [13] algorithm to obtain 20 different clusterings of each sample such that the number of clusters ranges from 1 to 20. The optimal number of clusters is determined as the number of clusters that maximizes *Calinski-Harabasz pseudo-f statistic* [14] (CH Index).

Following the clustering, a representative structure or alignment, called the *cluster centroid*, is computed for each of the identified clusters. The centroid of a cluster is the structure or alignment that has the smallest average distance of base pairs or

aligned positions, respectively, to all structures or alignments in the cluster. The centroid of the most populated cluster is utilized as estimated structure or alignment and the accuracy of these estimates is evaluated.

5. RESULTS

The structure prediction performance of joint sampling was compared to the single sequence sampling method that samples the Boltzmann distribution of structures of an RNA sequence [5] and the alignment prediction performance of joint sampling is compared to a sequence alignment hidden Markov model [15, 12]. All methods are evaluated on datasets containing 2000 randomly chosen pairs of tRNA sequences from the Sprinzl tRNA Database [16], 2000 randomly chosen pairs of 5S RNA sequences from the 5S Ribosomal RNA Database [17] and 40 randomly chosen pairs of RNase P sequences from the RNase P Database [18]. The cluster analysis was performed on the samples of 1000 structures and 1000 sequence alignments generated by joint sampling and single sequence sampling. This analysis identified on average 5.45 clusters for joint sampling and 4.31 clusters for single sequence sampling. The centroid structure or centroid alignment of the most populated cluster identified for each methods are evaluated in terms of accuracy.

The accuracy of predicted structures and alignments are reported in terms of sensitivity and positive predictive value (PPV). The sensitivity of structure prediction is the ratio of correctly predicted base pairs to the total number of base pairs in the known structure. The PPV is the ratio of number of correctly predicted base pairs to the total number of predicted base pairs. The sensitivity and PPV for prediction of alignments are defined similarly in terms of the number of correctly predicted aligned positions in the predicted alignment and number of aligned positions in the known alignment. Table 1 shows that the centroid structure of the most populated cluster for the joint sampling method is a more accurate predictor of secondary structure than its counterpart obtained via single sequence sampling. Results in Table 2 indicate that the centroid alignment of the most populated cluster from the joint sampling is more accurate than a hidden Markov model based sequence alignment prediction using sequence data alone without utilizing any structure considerations [15].

	Joint Sampling		Single Sampling	
	Sens.	PPV	Sens.	PPV
tRNA	0.808	0.862	0.716	0.683
5S RNA	0.737	0.766	0.642	0.609
RNase P	0.631	0.746	0.613	0.644

Table 1: Prediction accuracy of most populated cluster centroid structures in the sample sets generated by joint sampling and single sequence sampling.

	Joint Sampling		HMM	
	Sens.	PPV	Sens.	PPV
tRNA	0.857	0.856	0.794	0.787
5S RNA	0.940	0.946	0.906	0.902
RNase P	0.744	0.720	0.743	0.703

Table 2: Prediction accuracy of most populated cluster centroid alignments generated by joint sampling and the ML sequence alignment computed by sequence alignment hidden Markov model.

The joint sampling method is more computationally involved than single sequence sampling and the HMM based sequence alignment. The computation times for these algorithms are compared in [11]. The results in [11] also include a more exten-sive benchmarking of the algorithm performance compared with other multisequence RNA structural alignment algorithms.

6. CONCLUSION AND DISCUSSION

The improved structure prediction accuracy of joint sampling over the single sequence sampling demonstrates the advantage of utilizing multiple sequences in the sampling. In addition, the joint sampling utilizes structure commonality constraints in alignment prediction and offers improvements over the sequence alignment methods that utilize sequence data alone, for instance the sequence alignment hidden Markov model.

It is worth noting the centroids of few clusters identified by cluster analysis can be as alternative predictions for the structure of the sequence. The alternatives are valuable if there is experimental data to guide the choice of the correct conformation.

References

[1] J. D. Watson, T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick, *Molecular Biology of the Gene*, Pearson Education, Benjamin Cummings, San Franciso, CA, fifth edition, 2004.

[2] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Hum. Mol. Genet.*, vol. 15, no. 1, pp. 17–29, 2006.

[3] R. R. Gutell, J. C. Lee, and J. J. Cannone, "The accuracy of ribosomal RNA comparative structure models," *Curr. Opin. in Struct. Biol.*, vol. 12, pp. 301–310, 2002.

[4] D. H. Mathews and D. H. Turner, "Prediction of RNA secondary structure by free energy minimization," *Curr Opin Struct Biol*, vol. 16, no. 3, pp. 270–278, 2006.

[5] Y. Ding and C. E. Lawrence, "A statistical sampling algorithm for RNA secondary structure prediction," *Nucleic Acids Res.*, vol. 31, no. 24, pp. 7280–7301, 2003.

[6] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105 – 1119, November 1988.

[7] Y. Ding, C. Y. Chan, and C. E. Lawrence, "RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble," *RNA*, vol. 11, pp. 1157–1166, 2005.

[8] X. Xu, Y. Ji, and G. D. Stormo, "RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment," *Bioinformatics*, vol. 23, pp. 1883–1891, 2007.

[9] A. O. Harmanci, G. Sharma, and D. H. Mathews, "PARTS: Probabilistic alignment for RNA joinT secondary structure prediction," *Nucleic Acids Res.*, vol. 36, no. 7, pp. 2406–2417, April 2008.

[10] A. O. Harmanci, G. Sharma, and D. H. Mathews, "Probabilistic structural alignment of RNA sequences," in *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, Las Vegas, Nevada, Apr. 2008, pp. 645–648.

[11] A. O. Harmanci, G. Sharma, and D. H. Mathews, "Stochastic Sampling of the RNA Structural Alignment Space," *Nucleic Acids Research*, January 2009, Accepted for publication.

[12] A. O. Harmanci, G. Sharma, and D. H. Mathews, "Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign," *BMC Bioinformatics*, vol. 8, no. 130, April 2007.

[13] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data; An Introduction to Cluster Analysis*, Wiley, New York, NY, 2005.

[14] R. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1 – 27, 1974.

[15] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1999.

[16] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg, "Compilation of tRNA sequences and sequences of tRNA genes," *Nucleic Acids Res.*, vol. 26, pp. 148–153, 1998.

[17] M. Szymanski, M. Z. Barciszewska, J. Barciszewski, and V. A. Erdmann, "5S ribosomal RNA database Y2K," *Nucleic Acids Res.*, vol. 28, pp. 166–167, 2000.

[18] J. Brown, "The Ribonuclease P database," *Nucleic Acids Res.*, vol. 27, pp. 314, January 1999.